# A Genetics-based Machine Learning Approach to Knowledge Discovery in Clinical Data

**John H. Holmes, MS**
**College of Information Science and Technology, Drexel University**
**and**
**Center for Clinical Epidemiology and Biostatistics**
**University of Pennsylvania School of Medicine**
**Philadelphia, PA 19104**

This investigation focused on an evolutionary approach to machine learning, the *genetics-based classifier system*, and specifically its use in discovering knowledge structures in clinical data. In general, classifier systems operate on a knowledge base by means of evaluation, credit assignment, and discovery. These functions are handled, respectively, by three components: performance, reinforcement, and a genetic algorithm.

## METHODS

An object-oriented version of NEWBOOLE[1] (BOOLE++) was used in this study. The data used for this investigation were taken from the Wisconsin Breast Cancer Database[2] using a split-sample training-testing scheme. As the system was trained, its learning performance was evaluated by calculating the area under the receiver-operator characteristic (ROC) curve, as well as the proportion of unclassifiable cases (the Indeterminant Rate, or IR).

After BOOLE++ was trained, classification performance was evaluated by presenting each case in the testing set to the system. As was done during the training phase, the area under the ROC curve and the IR were calculated.

A decision rule was constructed by means of logistic regression (LR) analysis performed on the training set to derive a decision rule for classification. ROC curves were plotted for BOOLE++ and LR, and compared for significant difference in area.

## RESULTS

During training, BOOLE++ attained an area under the ROC curve of >0.95 within the first 100 iterations. The IR fell to <0.05 within 1,700 iterations, and by 26,000 iterations, the system was able to classify virtually all cases in the training set (IR<0.001).

After training, the classifier population was examined for the existence of knowledge structures. Of the 1,000 classifiers, a total of 339 (33.9%) were unique. Classifiers advocating positive classifications decoded to significantly higher total scores ($p<0.0001$). Only two pairs of classifiers (0.4%) advocated contradictory decisions.

The system exhibited excellent performance in classifying the cases in the testing set. The area under the ROC curve was 0.926 (SE=0.019), which compared well with that obtained by LR at a probability threshold of 0.90 (0.903, SE=0.029). No significant difference was found between the areas for BOOLE++ and LR ($p=0.175$).

## CONCLUSION

This study demonstrated the use of a genetics-based classifier system in discovering knowledge in clinical data. Learning these data occurred quickly. This was demonstrated in the nearly instantaneous (within 100 iterations) increase in the area under the ROC curve to nearly 1.00, and the fall of the IR to less than 5% within the first 1,700 iterations. The performance of BOOLE++ in classifying novel cases compared very well with that of a decision rule derived by logistic regression.

## REFERENCES

1. Bonelli P, Parodi A, Sen S, Wilson S. NEWBOOLE: A fast GBML system. In Porter B and Mooney R (eds.) Machine Learning: *Proceedings of the Seventh International Conference*, 1990 June 21, Austin, Texas. Morgan Kaufmann Publishers, San Mateo, CA, 1990, 153-159.
2. Murphy PM and Aha DW. UCI Repository of Machine learning Databases [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine, CA, 1992.

0195—4210/96/$5.00 